

# Stata Library

## Replicate Weights

### The what and why

The short answer to "what and why" is that replicate weights are a series of variables that contain the information necessary for correctly computing standard errors (using the replicate weight method) the standard errors of point estimates when analyzing survey data. Before we get into the specifics of what replicate weights are and how they are created, we need to know why they are needed in the first place. To understand that, we need to step back and look at how survey data is different from the analysis of data collected in other ways, e.g., experiments, quasi-experiments.

When we speak of survey data, we mean data that have been collected from subjects who were chosen based on a sampling plan. This is extremely important, because by using it, we have violated one of the assumptions of the statistical formulas used to calculate the statistics. The statistics described in most statistics texts assume that the data are collected based on a simple random sample of the elements of the survey research, this is almost never the case. Why? Because in most situations, it is too impractical and/or too expensive to collect data in a simple random sample. Because the SRS assumption has been violated, corrections to the calculation of the statistics are needed. There are two possible ways of making this correction. One way is called a Taylor series linearization method, and the other is called the replicate weight method. Before we can explain what the replicate weights are, we need to first understand some common elements found in many survey data sets (data sets). These elements are used in the Taylor series linearization method.

There are several elements that are unique to survey data that are necessary for correctly computing statistics based on the data. Each of these elements are variables that you will likely find in the data set. They are the probability weight (AKA sampling weight, pweight) variable, the PSU (primary sampling unit) variable, the stratification (AKA strata) variable, and the FPC (finite population correction) variable.

### Common elements of survey data sets

Most people do not conduct their own surveys with sampling designs. Rather, they use survey data that some agency or company collected and made available to the public. The documentation must be read carefully to find out what kind of sampling design was used to collect the data. This is important because many of the estimates and standard errors are calculated differently for the different sampling designs. Hence, if you misinterpret the sampling design, the point estimates and standard errors will likely be wrong.

Below are some common features of many sampling designs.

**Weights:** There are many types of weights that can be associated with a survey. Perhaps the most common is the sampling weight, sometimes called the probability weight, which is used to denote the inverse of the probability of being included in the sample due to the sampling design (except for the PSU, see below). The probability weight is calculated as  $N/n$ , where  $N$  = the number of elements in the population and  $n$  = the number of elements in the sample. For example, if a population has 10 elements and 3 are sampled at random with replacement, then the probability weight would be  $10/3$ . In a two-stage design, the probability weight is calculated as  $f_1 f_2$ , which means that the inverse of the sampling fraction for the first stage is multiplied by the inverse of the sampling fraction for the second stage. Under many sampling plans, the sum of the probability weights will equal the population size. For more information on weights, please see our [FAQ: What types of weights do SAS, Stata and SPSS support?](#)

**PSU:** This is the **primary sampling unit**. This is the first unit that is sampled in the design. For example, school districts from California may be sampled, then schools within districts may be sampled. The school district would be the PSU. If states from the US were sampled, and then schools within each state, and then schools from within each district, then states would be the PSU. One does not need to use the same sampling unit across levels of sampling. For example, probability-proportional-to-size sampling may be used at level 1 (to select states), while cluster sampling may be used at level 2 (to select school districts). In the case of a simple random sample, the PSUs and the elementary units are the same.

**Strata:** Stratification is a method of breaking up the population into different groups, often by demographic variables such as gender, race, or age. Once these groups have been defined, one samples from each group as if it were independent of all of the other groups. For example, if a sample is stratified on gender, men and women would be sampled independent of one another. This means that the probability weights for men will likely differ from the probability weights for the women. In most cases, you need to have two or more PSUs in each stratum. The purpose of stratification is to increase the precision of the estimates, and stratification works most effectively when the variance of the dependent variable is smaller within the strata than in the sample as a whole.

**FPC:** This is the **finite population correction**. This is used when the sampling fraction, the number of elements or respondents sampled relative to the population, becomes large. The FPC is used in the calculation of the standard error of the estimate. If the value of the FPC is close to 1, its impact is small and can be safely ignored. In some survey data analysis programs, such as SUDAAN, this information will be needed if you specify that the data were collected without replacement (see below for a definition of "without replacement"). The formula for calculating the FPC is  $\sqrt{(N-n)/(N+1)}$ , where  $N$  is the number of elements in the population and  $n$  is the number of elements in the sample. To see the impact of the FPC for samples of various sizes, suppose that you had a population of 10,000 elements.

Sample size (n)	FPC
1	1.0000
10	.9995
100	.9950
500	.9747
1000	.9487
5000	.7071
9000	.3162

### Sampling with and without replacement

Most samples collected in the real world are collected "without replacement". This means that once a respondent has been selected to be in the sample, that particular respondent cannot be selected again to be in the sample. Many of the calculations change depending on whether the sample is collected with or without replacement. Hence, programs like SUDAAN request that you specify if a survey sampling design was with or without replacement, and an FPC is used if sampling without replacement is used, even if the value of the FPC is very close to one.

## The why and how

Until recently, one needed to use special software (such as SUDAAN or WesVar) to correctly analyze survey data. Today, commonly used SAS, Stata and SPSS have procedures specially designed to handle the features of survey data. No matter which package is used, one still needs the probability weight, PSU, strata and FPC, if one is needed. The Taylor series linearization method of correcting the standard errors was the use of replicate methods mostly for computational purposes: It took less computing power to use the Taylor series. Back when computing was a major concern, this method became popular. However, a problem with this method arose (here is where we get to the replicate weight part!). In some number of respondents in a particular PSU was small, and people could start to figure out who the respondent was, even though no identifying information was contained in the data set. For a little example of how this works, suppose that we have a survey that is stratified on gender and race, and a few cities in Southern California. In some of these cities, there may be very few individuals in a particular strata, such as female Alaska Natives. If the survey figures out which PSU number corresponds to a particular city, the user may discover that there are only two female Alaska Natives in that strata. Perhaps other information in the survey, say age, can be used to determine exactly who the respondent is. Now the answers to the survey, supposed to be confidential, are no longer confidential. One way to avoid this problem is to not release data on strata that have fewer than a certain number of respondents in it. However, this can lead to misleading results because not all of the strata are being included in the analysis. Another solution is to use replicate weights. Because replicate weights are a series of many variables (often between 50 and 100) and their values are based on information provided to the user of the survey data set, it is nearly impossible for the user to figure out the identity of a given respondent. Note that when the jackknife method is used, the PSU and strata variables are not included in the data set. However, the probability weight will be included, and the probability weight and the replicate weights must be used for the correct calculation of the point estimate and its standard error.

There are several ways to create replicate weights. However, they are all based on a similar underlying logic. The sample is broken up into several called replicates. Next, the estimate of interest is calculated from both the full sample and from each replicate. Finally, the differences between the full sample and each of the replicates is used to determine the variance, i.e., the standard error, around the estimate. Different methods of creating the subsamples yield the different types of replicate weights. The different types of replicate weights include balanced repeated replication (JK-1, JK-2 and JK-n) and successive differences. The choice of what kind of replicate weight to create is determined by the type of sampling design that was used to collect the data, in particular, whether or not stratification was used and how many PSUs were in each strata. If stratification was used, the appropriate replicate weight method would be jackknife delete-1. If stratification was used and there were exactly two PSUs per strata, (or BRR with Fay's correction) or jackknife delete-2 could be used. If there were more than two PSUs per strata, jackknife delete-n would be appropriate. For more complete and extremely readable treatment of BRR and the various types of jackknife replicate weights, please see the WesVar manual, [see](#) for more information on successive difference replicate weights, please see [Fay and Train \(1995\)](#).

Besides protecting the privacy of survey respondents, the replicate weight method has other advantages. One is that the replicate weights contain information other than just about the strata and PSUs. Many surveys have corrections to the probability weight to account for nonresponse and/or raking to known totals, such as current Census figures. The effects of these adjustments can be incorporated into the replicate weights. There are some disadvantages to the replicate weight method. One is seen in extremely large data sets that have a huge number of replicates. In such instances, limitations of the software or computer could make the computation time extremely long or not possible. Another disadvantage is with the calculation of nonlinear statistics, such as ratios and quantiles. If the number of strata is small, there is a possibility of bias.

One last note about replicate weights. When specifying them in a program, you have to know by which method the replicate weights were created. If the estimates will be inaccurate if you "tell" the program that you have JK-1 replicate weights when in fact the replicates were formed using BRR. If replicate weights are provided as part of the data set, the documentation will tell you how the replicates were formed. This information can often be found in a section on the calculation of standard errors.

## Creating replicate weights

On rare occasions, one may need to create replicate weights for a survey data set. Several programs can be used for this. WesVar will create replicate weights, and there is a Stata .ado program by Nicholas Winters called **svr** (from the Stata command line, type **findit svr** to find and download it). Within this program is a command called **survwgt** that will create brr, jk1, jk2 and jkn replicate weights. A general introduction to the replicate weights (and Taylor series) can be found in chapter 4 of Analyzing Complex Survey Data by Eun Sul Lee, Ronald N. Fithian, and Ronald J. Lorim. The mathematical formulas on which the replicate weights are based can be found in many texts, including the WesVar 4 manual, which is online at [www.westat.com/Westat/pdf/wesvar/WV\\_4-3\\_Manual.pdf](http://www.westat.com/Westat/pdf/wesvar/WV_4-3_Manual.pdf). Documentation and a bibliography can be found at [http://www.westat.com/Westat/information\\_systems/WesVar/wesvar\\_documentation.cfm](http://www.westat.com/Westat/information_systems/WesVar/wesvar_documentation.cfm). Of course, Introduction to Variance Estimation by Kirk M. Wolter is the classic in the field.

## Using replicate weights in Stata

Now that we have a general idea about what replicate weights are and why they need to be used, it is time to use them. For our examples, we will use the adult CHIS data set (please see <http://www.chis.ucla.edu/>). The California Health Interview Survey (CHIS) is divided into several data sets, the "adult" data set. In the adult CHIS data set, there are 80 replicate weights created using the jackknife method (technically, the jackknife method). We will use these and the final pweight, called **rakedw0**, in our **svyset** command. In addition to specifying the probability weight, we also need to provide the jackweight adjustment multiplier, which for this data set, is 1. This information is found in the same package documentation that indicates how the replicate weights were created. If the type of replicate weights was BRR instead of jackknife, we would use a Fay's adjustment.

NOTE: The use of replicate weights is a new feature in Stata 9. The commands below will not work in earlier versions of Stata.

```
use "D:\temp\adult_chis.dta", clear
svyset [pw = rakedw0], jkrw(rakedw1 - rakedw80, multiplier(1))
```

Now that we have told Stata about the features of our data set, let's make sure that we have done this correctly. We can use the **svydes** command to check this. You will notice that, at the bottom of the output, there appears to be only one stratum and only one observation per unit (PSU). This is

information for the stratification and PSUs is contained in the replicate weights and therefore is not shown in that table.

## svydes

Survey: Describing stage 1 sampling units

```
pweight: rakedw0
VCE: linearized
jkrweight: rakedw1 rakedw2 rakedw3 rakedw4 rakedw5 rakedw6 rakedw7 rakedw8 rakedw9
rakedw12 rakedw13 rakedw14 rakedw15 rakedw16 rakedw17 rakedw18 rakedw19
rakedw22 rakedw23 rakedw24 rakedw25 rakedw26 rakedw27 rakedw28 rakedw29
rakedw32 rakedw33 rakedw34 rakedw35 rakedw36 rakedw37 rakedw38 rakedw39
rakedw42 rakedw43 rakedw44 rakedw45 rakedw46 rakedw47 rakedw48 rakedw49
rakedw52 rakedw53 rakedw54 rakedw55 rakedw56 rakedw57 rakedw58 rakedw59
rakedw62 rakedw63 rakedw64 rakedw65 rakedw66 rakedw67 rakedw68 rakedw69
rakedw72 rakedw73 rakedw74 rakedw75 rakedw76 rakedw77 rakedw78 rakedw79
Strata 1: <one>
SU 1: <observations>
FPC 1: <zero>
```

#Obs per Unit					
Stratum	#Units	#Obs	min	mean	max
1	55428	55428	1	1.0	1
1	55428	55428	1	1.0	1

Next, we will run a simple regression example using **ae13** as the response (dependent) variable and **ae14** as the predictor (independent) variable. Both variables were chosen at random. Although the command will run (and run faster) without the jackknife option after the **svy**, you will get linear errors instead of the jackknife standard error. This jackknife standard error matches the standard errors produced by both SUDAAN and W.

## svy jackknife: regress ae13 ae14

```
Jackknife replications (80)
-----+----- 1 -----+----- 2 -----+----- 3 -----+----- 4 -----+----- 5
..... 50
.....
```

Survey: Linear regression

```
Number of strata   =          1
Number of obs      =       55428
Population size    =    23847415
Replications       =          80
Design df         =          79
F(      1,      79) =       648.74
Prob > F           =       0.0000
R-squared          =       0.2419
```

	ae13	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]
	ae14	.3356038	.0131763	25.47	0.000	.3093772 .3618305
	_cons	1.884704	.0123061	153.15	0.000	1.86021 1.909199

## Sample SUDAAN setups are helpful

Because SUDAAN has been able to handle replicate weights much longer than Stata has, the official documentation for a survey may include setup for SUDAAN but not for Stata, although you might find some Stata examples on the web. Don't brush by the SUDAAN example think useless to you as a Stata user; rather, it is often the easiest way to get all of the information that you need for your **svyset** command. All survey analysis programs need to have the same information: the probability weight, type of replicate weight that is to be used, the names of the replicate variables, and the adjustment factor. These elements are needed regardless of the type of sampling plan that was used. In SUDAAN, the information will be listed on the **weight** statement. The type of weight will be listed in the **design** option on the **proc** statement. The names of the replicate weights will be found on the **jackwgt**s statement for jackknife replicate weights or on the **repwgt** statement for BRR replicate weights. The adjustment

the same statement - **adjjack** for jackknife replicate weights and **adjfay** for BRR replicate weights.

## A little note about pseudo-strata and pseudo-PSUs

Some modern data sets are being released with pseudo-strata and pseudo-PSUs. These can be used in a Taylor series linearization just as pseudo counterparts would be. These elements are "pseudo" in the sense that they have been modified so that, while the point estimates and errors are estimated correctly, users of the data set are unable to use the strata and PSU information to figure out who individual responder results obtained using these pseudo elements may differ more from the results obtained using the replicate weights than the results from using pseudo elements. You may find that wider confidence intervals are obtained when using the pseudo-strata and pseudo-PSUs than when using weights, if both are available in the data set.

[How to cite this page](#)

[Report an error on this page or](#)

The content of this web site should not be construed as an endorsement of any particular web site, book, or software product by the University of Michigan.